Correlation-based biological networks

Won-Min Song[a], Tomaso Aste[a] and T. Di Matteo[a]

[a]Department of Applied Mathematics,
Research School of Physical Science and Engineering,
The Australian National University, 0200 Canberra, ACT, Australia.

ABSTRACT

We construct a correlation-based biological network from a data set containing temporal expressions of 517 fibroblast tissue genes at transcription level. Four relevant and meaningful connected subgraphs of the network, namely: minimal spanning tree, maximal spanning tree, combined graph of minimal and maximal trees, and planar maximally filtered graph are extracted and the subgraphs' geometrical and topological properties are explored by computing relevant statistical quantities at local and global level. The results show that the subgraphs are extracting relevant information from the data set by retaining high correlation coefficients. The design principle of the underlying biological functions is reflected in the topology of the graphs.

1. INTRODUCTION

Real complex systems such as acquaintance networks, World-Wide-Web and metabolisms are made of many interacting elements connected through a system of links which possesses common properties such as scale-free degree distribution and small-worldness (small average path length) despite of the differences in their nature.[1] The use of network approaches to study real complex systems has enabled scientists to gather insights into biological networks such as metabolic pathway networks[2] and protein networks.[3] In particular, large-scale data for genomes of simple organisms such as *E.coli* and *S.cerevisae* have accumulated sufficient enough information to retrieve the global structure and properties of the networks regulating many genes interacting via complicated chemical reactions.[4,5]

In this paper, we approach the study of biological data sets by taking an opposite direction. Given no prior information on the underlying biological or chemical links, we analyze the correlations in the temporal expressions to construct connected graphs representing the most meaningful and relevant information of the data set. We show that, by doing so, one can partially retrieve geometrical and topological properties of the underlying biological network.

Four different graphs are constructed in order to perform this analysis. In particular, two trees: minimal spanning tree (Γ_{min}) and maximal spanning tree (Γ_{max}) are constructed as the most skeletal graphs to extract the relevant information. These trees, Γ_{min} and Γ_{max}, are then combined to give Γ_{trees}. Furthermore, the planar maximally filtered graph (Γ_P) is constructed to provide richer information by selectively accepting highly weighted edges which contain meaningful information with less strict constraints than the minimal spanning tree. In this paper, Γ_P is applied to a biological data set for the first time. The technique has already been shown to successfully retrieve relevant information in a financial data set.[6] Though the financial data set is different in nature in comparison to the biological data set, we show that the generality in network approach enables one to unveil the most core information in the biological data set as well. In Section 2 the biological data set is described, in Section 3 the filtered graphs algorithms are constructed and several statistical quantities computed in the graphs are introduced. The results are reported in Section 4 and conclusions in Section 5.

Further author information:
Won-Min Song: E-mail: wms110@rsphysse.anu.edu.au, Telephone: +61 422 185 579

Complex Systems II, edited by Derek Abbott, Tomaso Aste, Murray Batchelor, Robert Dewar,
Tiziana Di Matteo, Tony Guttmann, Proc. of SPIE Vol. 6802, 680212, (2008)
0277-786X/08/$18 · doi: 10.1117/12.759252

2. THE DATA SET

Fibroblast tissues are known to play important roles in physiology of wound healing. They participate in cell cycle and proliferation, angiogenesis, re-epithelialization, cytoskeletal reorganization etc.[7] Vishwanath *et al.*[7] have extracted from about 8,600 distinct human fibroblast genes a subset of 517 human fibroblast genes which showed significant responses under stimulation by Fetal Bovine Serum (FBS). The serum stimulation signifies a signal to notify detection of an wound within a human body.[7]

The responses were observed by using a microarray technique. The genes were first induced to enter quiescent state by depriving serum then stimulated by FBS. The responses were observed at 12 time points ranging from 15 min to 24 hrs. Including the first time point before the stimulation, the data set contains $N = 517$ genes' responses at 13 time points. Expression levels are observed relatively to the initial time point by normalizing with the expression level at the first time point. The entire information is encapsulated in the following 517×13 matrix.

$$\Lambda_{fibro} = \begin{pmatrix} \tilde{v}_1 \\ \vdots \\ \vdots \\ \tilde{v}_N \end{pmatrix}$$

where $\tilde{v}_i = [x_1 \ldots x_k \ldots x_{13}]$ is a row vector representing an expression profile of the genes.

3. METHODOLOGIES

3.1 Metric distance

The metric distance between any two expression profiles is computed by[8]

$$d_{ij} = \sqrt{2(1 - \rho_{ij})} \tag{1}$$

where $\rho_{ij} = Cor(\tilde{v}_i, \tilde{v}_j)$ is the correlation coefficient between the profiles i and j. Since the correlation coefficient measures similarity between two profiles, this can be taken as the similarity index so that similar profiles possess small distances accordingly. ***Two genes whose expression profiles obtain a high correlation coefficient ($\rho_{ij} \sim 1$), hence a small distance, are very likely to be involved in a same role in the physiology of wound healing.*** Conversely, two genes whose expression profiles obtain a highly negative correlation coefficient ($\rho_{ij} \sim -1$), hence a large distance, are very likely to be involved in two opposite roles.

3.2 Construction of skeletal correlation-based connected graphs

3.2.1 Necessity to construct skeletal graphs

Applying Eq. 1 to every pair of gene expression profiles yields to a complete graph, G_{com}, with 517 nodes and $517(517 - 1)/2 = 133,386$ undirected edges whose weights are assigned by the distances function or the correlation coefficients. Fig. 1 shows the probability distribution of correlation coefficients ρ_{ij} in the complete graph generated by the data set. The $P(\rho_{ij})$ shows that there are approximately as many edges whose ρ_{ij} are close to 1 or -1 as well as around 0. This means there is as much meaningful information as not meaningful information. Therefore, the relevant information must be 'distilled' so that an effective inference on the data set can be performed.

In this paper, the 'distilling process' takes place by constructing a connected skeletal subgraph of G_{com} enforcing topological constraint to control how skeletal the constructed subgraph is. Specifically, the following two types of topological constraints are employed:

1. Generate trees which span all the nodes (spanning trees) and the sum over the distances over the connected nodes is minima (Γ_{min}) or maximal (Γ_{max}).

2. Generate planar graphs (whose edges can be embedded on a spherical surface without edge crossings) with minimal sum of the distances over the connected nodes.

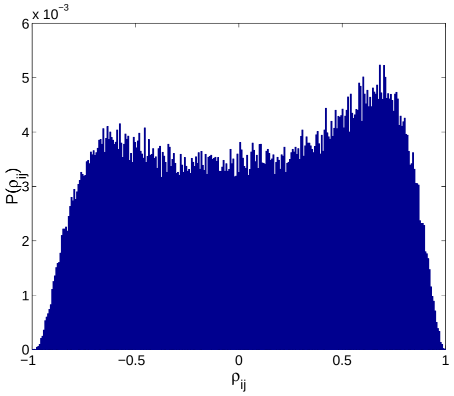

Figure 1. Probability distribution of correlation coefficients ρ_{ij}s for G_{com}.

3.2.2 Construction of a Minimal Spanning Tree

A spanning tree, T, is defined as a tree[*] which connects all the nodes in G_{com}. Minimal Spanning trees have been thoroughly studied in diverse fields such as invasion percolation[1] and information structured graphs.[9] In network theory, they have been suggested as skeletal subgraphs which are representative of the original graphs.[10,11] The Γ_{min} of G_{com} is constructed by using a greedy algorithm called Prim's algorithm [†]. The resulting tree possesses 516 edges connecting 517 nodes and is unique for the G_{com}.

3.2.3 Construction of a Maximal Spanning Tree

Edges with highly negative correlation coefficients are just as significant in terms of genetics as those with positive correlation coefficients. Indeed, gene expressions suppressing others by inducing itself or inducing others by suppressing itself are just as meaningful.[12] These will be referred as 'switching on and off' activities in the data set as the profiles involved in an edge with a high negative ρ_{ij} behave opposite to the other. The same algorithm for constructing Γ_{min} has been applied to construct a 'Maximal Spanning Tree', Γ_{max} containing edges with most negative ρ_{ij}s. The resulting tree possesses 516 edges connecting 517 nodes and is unique for the G_{com}.

3.2.4 Construction of a Planar Maximally Filtered Graph

Realizing a subgraph of G_{com} as a triangulation of a hyperbolic surface with genus[‡] g has been recognized as an effective tool for extracting relevant information of G_{com} with complexity controlled by g.[13] The algorithm we use connects edges from a ordered list from the lowest to the largest distances accepting the connection if and only if it does not cross another edge already embedded.[6] When $g = 0$, the resulting subgraph of G_{com} is called the Planar Maximally Filtered Graph of G_{com}, Γ_P, and it has $3(N-2) = 1,545$ edges where $N = 517$ is the number of genes involved. This method is here applied to biological data for the first time. The graph Γ_P includes Γ_{min} as a subgraph of its own.[6]

3.3 Explored quantities

The following quantities are computed to explore topological and geometrical properties of the extracted skeletal subgraphs.

1. *Weight distribution,* $P(\rho_{ij})$ is the probability distribution of the correlation coefficients ρ_{ij} in the extracted subgraphs. It is appropriate to choose ρ_{ij} as 'weight' of an edge because ρ_{ij} represents the likelihood of how closely two genes are related and ρ_{ij} is high when two genes are closely related so that it scales in the right direction with common intuition having high weight when an edge is important.

[*] A graph with no cycles.

[†] A greedy algorithm navigates with local information only. Γ_{min} is not obtained by the entire information of G_{com} at once.

[‡] That is, a spherical surface with g handles.

2. *Acceptance function, $A(\rho_{ij})$,* is the probability of an edge with weight ρ_{ij} in G_{com} to be involved in a subgraph.

3. *Degree distribution, $P(k)$,* is the probability distribution for the gene connectivity ('degree'). The degree distributions of the skeletal subgraphs, $P(k)$s, are explored to see how each subgraph gives different varieties of degrees. Genes involved in a large number of meaningful interactions are expected to have high degrees.

4. Node correlation, r: Assortative mix in each skeletal subgraph is measured by using the correlation coefficients:[14]

$$r = \frac{\langle k_i k_j \rangle - \langle k_i \rangle \langle k_j \rangle}{\sqrt{\langle (k_i - \langle k_i \rangle)^2 \rangle} \sqrt{\langle (k_j - \langle k_j \rangle)^2 \rangle}} \tag{2}$$

   where $k_i,\ k_j$ denote for degrees of nodes at both ends of an edge in the graph and the averages are over the whole set of edges in the graph.

5. Node strength as a function of degree, $s(k)$: Node strength of ith node is defined as:

$$s_i = \sum_j \rho_{ij}. \tag{3}$$

   Then $s(k)$ is defined as average of node strengths of nodes with degree k. This is useful for detecting correlation between node strength and degree. If there are no correlations, then $s(k) = Ak$ with $A = \langle \rho_{ij} \rangle$. Otherwise, $s(k) = Ak^{\alpha}$ with $A \neq \langle \rho_{ij} \rangle$ and α not necessarily equal to 1.[15]

6. Disparity as a function of degree, $Y(k)$: Disparity of ith node is defined as:

$$Y_i = \sum_j \left[\frac{\rho_{ij}}{s_i} \right]^2. \tag{4}$$

   This measure is particularly useful for detecting presence of edges with dominant weight(s) at node. If all weights are comparable, then $Y_i \sim 1/k_i$. Conversely, if there is a dominant weight, then $Y_i \sim 1$.
   $Y(k)$ is then defined as average of disparities of nodes with degree k. This is useful for detecting homogeneity of weights within the whole skeletal subgraph. If the majority of edges have comparable weights, then, $Y(k) \sim 1/k$. If not, $Y(k) \sim 1$.[15]

4. RESULTS

Acceptance functions of the subgraphs, $A(\rho_{ij})$s, and assortative mix within each subgraphs are explored. A random data set containing 517 random walk profiles with each walk generated by a normalized Gaussian distribution[§] was artificially generated and analyzed in the same fashion as the real data set. In order to establish if the results obtained from the real data set are meaningful we have compared these results to the null case results by generating a random data set.

In Figs. 2, 3, 4 and 5 are reported the behaviors of the acceptance function calculated for $\Gamma_{min},\ \Gamma_{max},\ \Gamma_P$ for both the real data set and the random data set. The linear trends in the log-normal plots reveal that $A(\rho_{ij})$ is mainly characterized by an exponential behavior $A(\rho_{ij}) = a\exp(-\zeta\rho_{ij})$. In Table 1, we report the best fit values for the coefficients ζ. The fitting curves are also shown in the Figs. 2 to 5.

Distinctions between the real data set and artificial random data set were found in the exponent ζ. $A(\rho_{ij})$s of subgraphs from the real data set have higher exponents than the $A(\rho_{ij})$s of corresponding subgraphs from the random data set so that acceptance of edges in G_{com} of the real data set favors higher weight edges than that of the random data set. This indicates that the subgraphs from the real data contain more information than

[§] A Gaussian distribution with mean=0 and standard deviation=1.

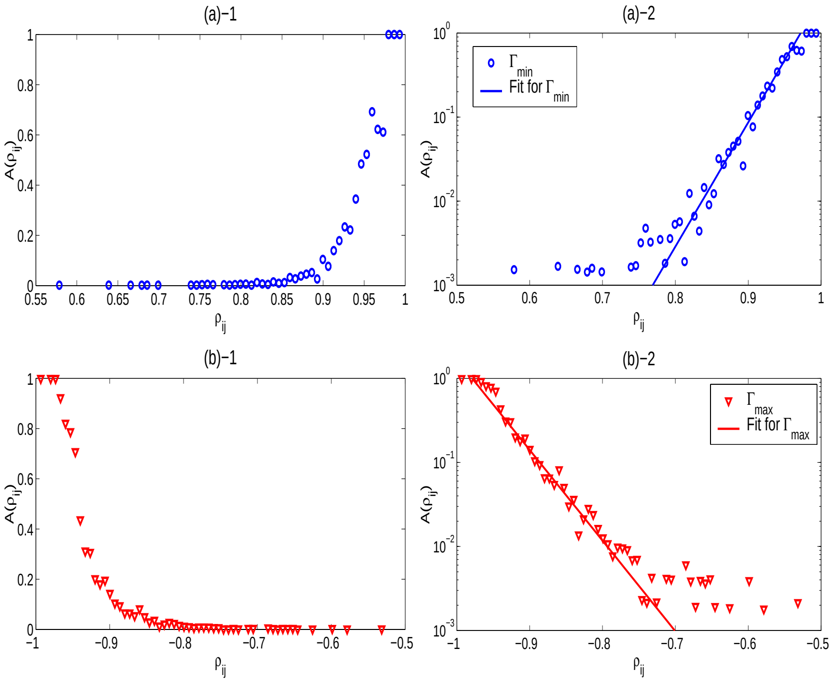

Figure 2. Probability of accepting an edge with weight ρ_{ij} in Γ_{min}, $A(\rho_{ij})$, has been plotted for the real data set in (a)-1. The $A(\rho_{ij})$ has been fitted with an exponential in (a)-2. Similarly, $A(\rho_{ij})$ of Γ_{max} of real data set is plotted in (b)-1, and the fitted exponential function is shown in (b)-2.

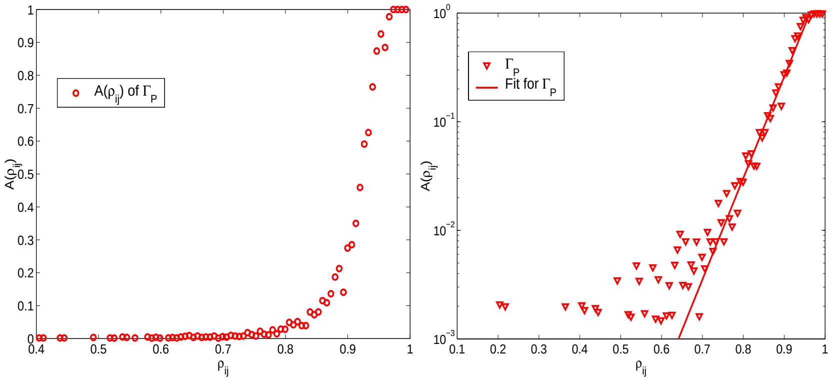

Figure 3. (a) Probability of accepting an edge with ρ_{ij} in G_{com}, $A(\rho_{ij})$, in Γ_P for the real data set. The corresponding exponential fit is displayed in (b).

Real data set	ζ
Γ_{min}	34.0245
Γ_{max}	24.8413
Γ_P	21.4915
Random data set	ζ
Γ_{min}	11.1697
Γ_{max}	11.8190
Γ_P	8.6633

Table 1. Fitted exponent for $A(\rho_{ij})$s in the subgraphs Γ_{min}, Γ_{max}, Γ_P from both data sets.

independent random walks.

The exponential for any subgraphs favors high weight edges. Thus, the subgraphs have extracted from the both data sets highly correlated and anti-correlated profiles. $A(\rho_{ij})$s of a Γ_P accepts more edges than the Γ_{min} of the same data set indeed it has Γ_{min} as its own subgraph[6] and it is especially generous in accepting more edges in the tail region than the Γ_{min}. This is reflected in the smaller value of the exponent and in the fact that Γ_P is decaying slower than exponential in the tail. This allows Γ_P to include more diverse disparities and node

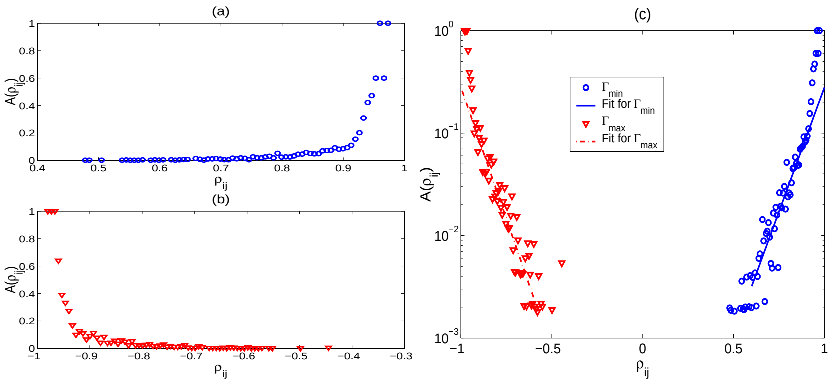

Figure 4. Probability of accepting an edge with ρ_{ij} in G_{com}, $A(\rho_{ij})$, in (a) Γ_{min} and (b) Γ_{max} of the random data set. The exponential fits are displayed in (c) (Red plots for Γ_{max}, Blue plots for Γ_{min}).

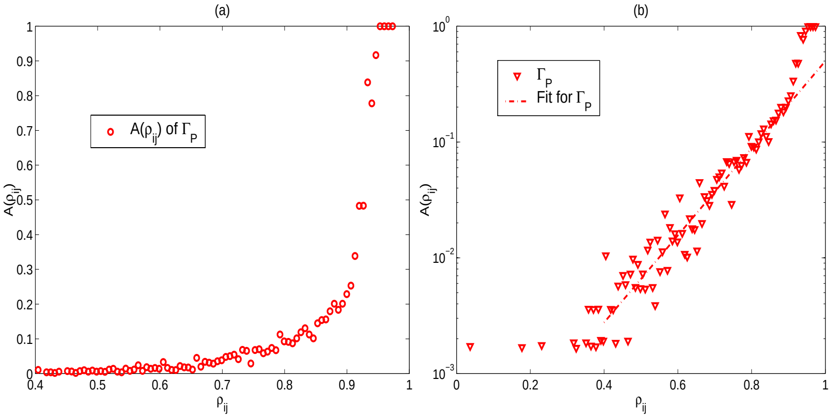

Figure 5. Probability of accepting an edge with ρ_{ij} in G_{com}, $A(\rho_{ij})$, in Γ_P of the random data set is plotted in (a). The exponential fit is displayed in (b).

strengths than a Γ_{min}.
One can note from Table 1 that, the fitted exponents for Γ_{min} and Γ_{max} for the random data set have values close to each other with opposite signs. This is a consequence of the symmetry about $\rho_{ij} = 0$ so that finding positively correlated profiles is equivalent to finding negatively correlated profiles. This yields to symmetric Γ_{min} and Γ_{max} in terms of $A(\rho_{ij})$. Such symmetry is not observed in the real data set. The fitted exponent in Table 1 shows that $A(\rho_{ij})$ for Γ_{max} has fatter tail than that of Γ_{min}. This suggests there is some meaningful restriction which limits Γ_{max} to have less high weighted edges than Γ_{min}.

4.1 Common Results obtained from Γ_P and Γ_{min}

Though Γ_P has richer information than Γ_{min}, we have found that they also share some common properties.

4.1.1 Homogeneity in $P(\rho_{ij})$ and $P(k_i)$

The probability distributions of the $P(\rho_{ij})$s of both graphs show exponential decay for both real and random data sets (Figs. 6, 7). The distributions also have high means close to $|\rho_{ij}| = 1$. This implies that the subgraphs have the majority of edge weights concentrated close to $|\rho_{ij}| = 1$ so that each subgraph has a homogeneous weight distribution.
The degree distributions in the subgraphs, $P(k_i)$s, have also exponential tails in both of Γ_{min} and Γ_P for real and random data sets (see Figs. 8, 9). The fitting of the cumulative distributions in a log-normal scale are repeated in Figs. 8 and 9. These plots also indicate a homogeneous degree distribution. The homogeneity in each subgraph is well reflected in behavior of $Y(k)$ which has been well fitted with $Y(k) \sim 1/k$ in Fig. 11.
The above results indicate that the real data set contains information on the underlying central interactions

which are rather homogeneous. Indeed, the data set is a collection of significant results obtained from a pre-filtered gene set from about 8,600 genes,[7] which would be likely to collect only central and strong interactions between the genes. The interactions filtered by the subgraphs are likely to be the strongest interactions among those present in the entire genome of a fibroblast tissue. This omits contributions of weak interactions which would contribute to the heterogeneity of the underlying network.

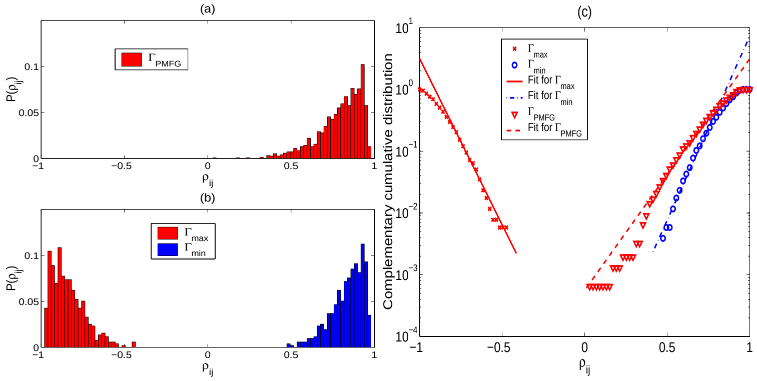

Figure 6. Probability distribution of correlation coefficients, $P(\rho_{ij})$, of (a) Γ_P (b) Γ_{min} and Γ_{max} for the random data set. (c) Complementary cumulative distributions of ρ_{ij} and fitted exponential distributions.

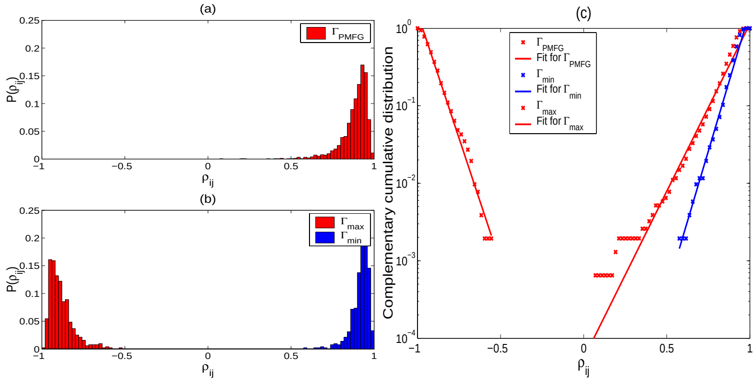

Figure 7. Probability distribution of correlation coefficients, $P(\rho_{ij})$, of (a) Γ_P (b) Γ_{min} and Γ_{max} for the real data set. (c) Complementary cumulative distributions of ρ_{ij} and fitted exponential distributions.

4.1.2 Correlation between s_i and k_i

Though the results are homogeneous, we see that the subgraphs have captured some distinct characteristics of the real data set. The fitting of $s(k)$ with a linear law $s(k) = ak$ gives values for the weight a (see Table 2) which are larger than the mean weight in both subgraphs and this indicates that a larger degree node obtains larger weight per edge than a smaller degree node. This suggests presence of hub nodes with strongest interactions with neighboring genes.

	Γ_{min}	Γ_{max}	Γ_P
Fitted coef.	$a_1 = 0.9476$	$a_2 = -0.8889$	$a_p = 0.9258$
$\langle\rho_{ij}\rangle$	0.9179	-0.8828	0.8825

Table 2. Linear functions have been fitted to node strength as a function of degree for various subgraphs and the fitted coefficients are displayed in the top row. Subscript '1' reserved for Γ_{min}, '2' for Γ_{max}, 'P' for Γ_P. Each coefficient is compared to the mean weight, $\langle\rho_{ij}\rangle$, within each subgraph which are listed in the bottom row.

4.1.3 Insignificant node correlation

Eq. 2 has been used to detect assortative mix within each subgraph. The amount of assortative mix in the trees of the random data set has been calculated to see if the trees from the real data set have any significant

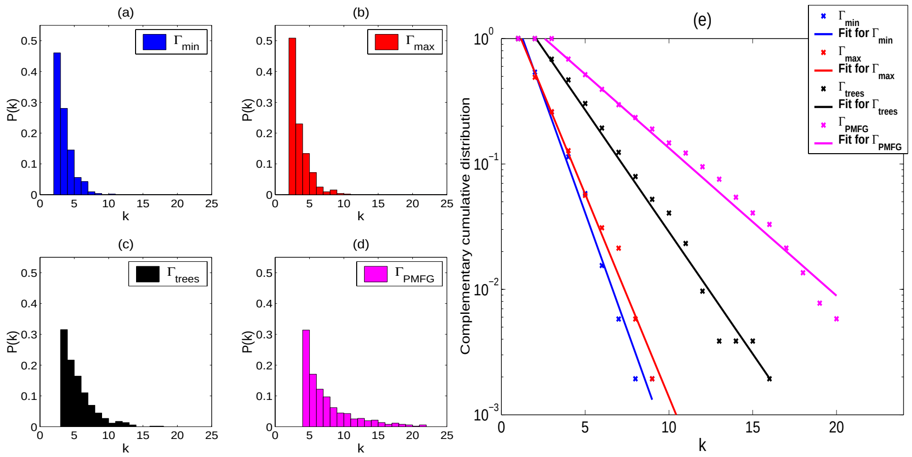

Figure 8. Probability distribution of degree, $P(k_i)$, of (a) Γ_{min} (b) Γ_{max} (c) Γ_{trees} and (d) Γ_P of the random data set. (e) Complementary cumulative distributions of k and fitted exponential distributions.

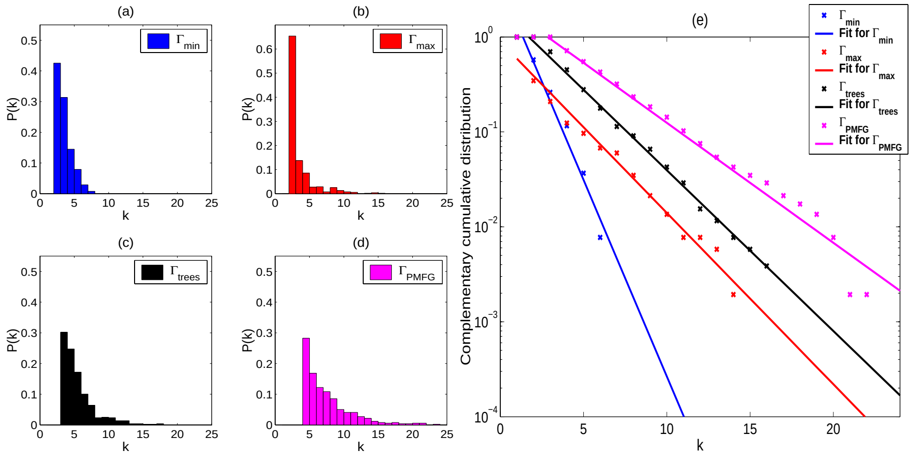

Figure 9. Probability distribution of degree, $P(k_i)$, of (a) Γ_{min} (b) Γ_{max} (c) Γ_{trees} and (d) Γ_P of the real data set. (e) Complementary cumulative distributions of k and fitted exponential distributions.

amount of assortative mix. Table 3 shows that Γ_{min} and Γ_P of the real data set do not obtain significant mix in comparison to those of the random graphs as they differ by only ~ 0.01. The only significant coefficient was detected in Γ_{max} which shows fairly large amount of disassortative mix in node degree for the real data set. This indicates that repressions (or inductions) of genes with high degrees in Γ_{max} take place efficiently to induce (or repress) the neighboring genes with low degrees.

	Random data set	Real data set
Γ_{min}	-0.1307	-0.1893
Γ_{max}	-0.1571	**-0.5072**
Γ_P	-0.0198	-0.0095

Table 3. Node correlation coefficients computed for Γ_{min}, Γ_{max} and Γ_P of the real data set (left column) and the random data set (right column). Eq. 2 has been used for the computation. Bold number is to emphasize significant degree mix within the corresponding subgraph.

4.2 Assortative mix of degrees between Γ_{min} and Γ_{max}

Both trees have been combined into a single subgraph, Γ_{trees}, to observe correlation between node degrees in Γ_{min} and in Γ_{max}. Since these two trees possess either all positive or negative edge weights, they have disjoint

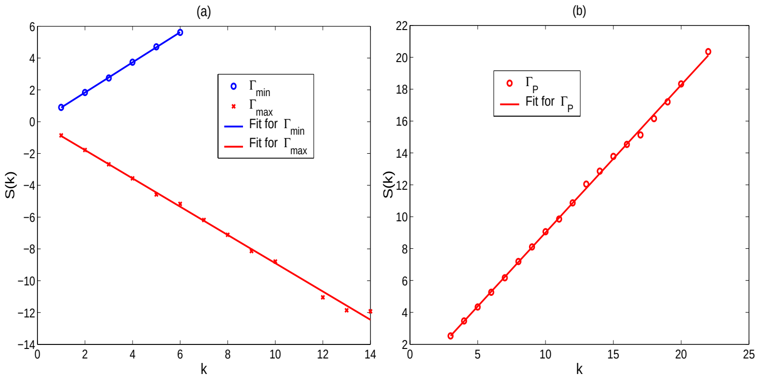

Figure 10. Node strength as a function of degree, $s(k)$, in Γ_{min} (blue plot) and Γ_{max} (red plot) of the real data set in (a). $s(k)$ for Γ_P of the real data set is plotted in (b).

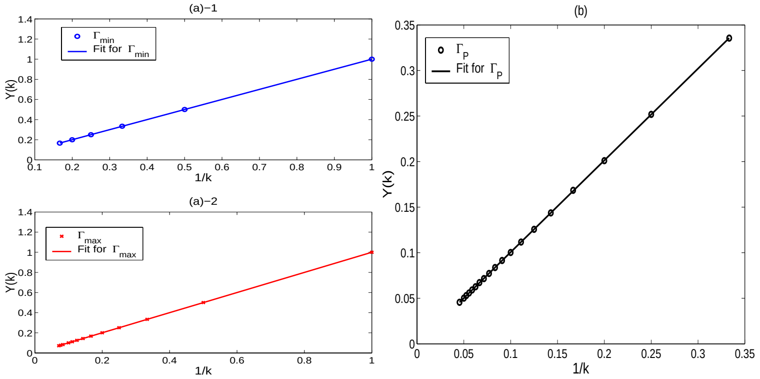

Figure 11. Disparity as a function of degree, $Y(k)$, has been plotted as a function of $1/k$. (a)-1 plots for Γ_{min} and (a)-2 for Γ_{max} of the real data set. (b) plots for Γ_P of the real data set.

edges sets sharing a common node set, so Γ_{trees} consists of $516 \times 2 = 1,032$ edges and 517 nodes. Denoting node degrees in Γ_{min} as k_i^{min} and those in Γ_{max} as k_j^{max}, two aspects of node degree correlation are investigated:

- Assortative mix at the nodes to detect any preference of a node with k_i^{min} in Γ_{min} to have a certain degree k_i^{max} in Γ_{max}. We denote this correlation coefficient as r_{node}.
- Assortative mix via the edges to detect any preference of edges to attach a node with k_i^{min} to some degree node k_j^{max} via edges of Γ_{min} or vice versa. We denote these correlation coefficients for assortative mix via edges in Γ_{min} as r_{edge}^{min} and via edges in Γ_{max} as r_{edge}^{max}.

These correlation coefficients have been obtained by computing the Pearson correlation coefficients for each case. The computation is based on the correlation function which measures the amount of departure from the null correlation by:

$$\langle k_i^{min} k_j^{max} \rangle - \langle k_i^{min} \rangle \langle k_j^{max} \rangle. \tag{5}$$

The correlation coefficients are normalized by the maximal value in each case, which represents the most assortative mix and is given by standard deviations:

$$r_{node} = \frac{\langle k_i^{min} k_i^{max} \rangle_{node} - \langle k_i^{min} \rangle_{node} \langle k_i^{max} \rangle_{node}}{\sqrt{\langle (k_i^{min} - \langle k_i^{min} \rangle)^2 \rangle_{node}} \sqrt{\langle (k_i^{max} - \langle k_i^{max} \rangle)^2 \rangle_{node}}} \tag{6}$$

$$r_{edge}^{min} = \frac{\langle k_i^{min} k_j^{max} \rangle_{edge}^{min} - \langle k_i^{min} \rangle_{edge}^{min} \langle k_j^{max} \rangle_{edge}^{min}}{\sqrt{\langle (k_i^{min} - \langle k_i^{min} \rangle_{edge}^{min})^2 \rangle_{edge}^{min}} \sqrt{\langle (k_j^{max} - \langle k_j^{max} \rangle_{edge}^{min})^2 \rangle_{edge}^{min}}} \tag{7}$$

$$r^{max}_{edge} = \frac{\langle k^{min}_i k^{max}_j \rangle^{max}_{edge} - \langle k^{min}_i \rangle^{max}_{edge} \langle k^{max}_j \rangle^{max}_{edge}}{\sqrt{\langle (k^{min}_i - \langle k^{min}_i \rangle^{max}_{edge})^2 \rangle^{max}_{edge}} \sqrt{\langle (k^{max}_j - \langle k^{max}_j \rangle^{max}_{edge})^2 \rangle^{max}_{edge}}}. \quad (8)$$

In terms of adjacency matrix, the means are computed by:

$$\begin{aligned} \langle f(k_i) \rangle_{node} &= \frac{1}{N} \sum_i f(k_i) \\ \langle f(k_i, k_j) \rangle^{min}_{edge} &= \frac{1}{2(N-1)} \sum_{i,j} a^{min}_{i,j} f(k_i, k_j) \\ \langle f(k_i, k_j) \rangle^{max}_{edge} &= \frac{1}{2(N-1)} \sum_{i,j} a^{max}_{i,j} f(k_i, k_j) \end{aligned}$$

where N is the number of nodes in both graphs, $a^{min}_{i,j}$ and $a^{max}_{i,j}$ are the adjacency matrices of Γ_{min} and Γ_{max}.

	Real data set	Random data set
r_{node}	0.0851	0.4818
r^{min}_{edge}	0.0127	-0.0020
r^{max}_{edge}	0.1306	0.0512

Table 4. Correlation coefficients to indicate any assortative mix between k^{min}_i and k^{max}_j in Γ_{trees}.

Table 4 lists the computed coefficients for both the real data set and the random data set. The differences are evident from r_{node} showing that the real data set has no significant correlation between k^{min}_i and k^{max}_j at nodes while the random data set has large correlation. This large r_{node} can be explained by the fact that a node which obtains a high degree in one tree is likely to obtain a high degree in the other tree since Γ_{min} and Γ_{max} are symmetric. It is because the random walk diffusion takes place symmetric in both directions. The low r_{node} in Γ_{trees} of the real data set can be interpreted as there is almost no relation between genetic role of a gene in Γ_{min} and another genetic role of the gene in Γ_{max} and vice versa.

The data for r^{max}_{edge} shows that significant assortative mix via edges from Γ_{max} is present in Γ_{trees} for the real data set. The significant r^{max}_{edge} indicates that high degree nodes in Γ_{max} tend to connect to high degree nodes in Γ_{min}. Recalling that the edges in Γ_{max} represent switching on or off activities within the genome of a fibroblast tissue in response to the serum stimulation, these results suggest that the switching processes take place on high degree nodes in Γ_{min} to effectively transmit the signal to activate genetic activities related to the physiology of wound healing.

5. CONCLUSION

There have been various attempts to construct biological networks from microarray data sets. Boolean networks,[16] differential equations,[17] and Bayesian networks[18] are general approaches usually taken in Bioinformatics. In this paper, without any prior knowledge on the genetics of the corresponding genome, the geometrical and topological properties of gene interactions have been detected by constructing different types of skeletal correlation-based networks.

The construction of subgraphs of G_{com} such as Γ_{trees} and Γ_P has been proven to be a valid instrument to capture some of the biological information contained in the expressions of genes in the data set despite of the large amount of filtering applied onto the information from the genome of interest. In particular, the less strict constraint to construct a Γ_P improved $A(\rho_{ij})$ to favor high weights as well as generous acceptance for lower weight edges in $\rho_{ij} \in [0.5, 0.85]$ respect to Γ_{min}.

It has been found that high degree nodes in Γ_{min} tend to have larger edge weights than low degree nodes, and high degree nodes in Γ_{max} tend to connect to high degree nodes in Γ_{min}. This suggests that the genetic transcription activities of genes from human fibroblast tissue respond to the detection of wound to efficiently

transmit the signal through highly efficient 'switches' (that is, high degree nodes in Γ_{max}) acting on high degree nodes in Γ_{min} which are very effective in activating several genetic activities.
The detected properties are not only meaningful topologically, but also verified in terms of known genes interactions in wound healing.

Acknowledgements

Many thanks to Mario Nicodemi and Massimo Pica Ciamarra for many useful discussions. This work was partially supported by the ARC Discovery Projects DP0344004 (2003), DP0558183 (2005) and COST P10"Physics of Risk" project.

REFERENCES

1. R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics* **74**, pp. 48–94, 2002.
2. H. Jeong, B. Tombor, R. Albert, Z. Oltvai, and A.-L. Barabási, "The large-scale organization of metabolic networks," *Nature* **407**, pp. 651–654, 2000.
3. H. Jeong, S. Mason, A.-L. Barabási, and Z. Oltavai, "Lethality and centrality in protein networks," *Nature* **411**, pp. 41–42, 2001.
4. R. Milo *et al*, "Network motifs: Simple building blocks of complex networks," *Science* **298**, pp. 824–827, 2002.
5. S. Wuchty, Z. Oltavai, and A.-L. Barabási, "Evolutionary conservation of motif constituents in the yeast protein interaction network," *Nature Genetics* **35**, pp. 176–179, 2003.
6. M. Tumminello, T. Aste, T. D. Matteo, and R. Mantegna, "A tool for filtering information in complex systems," *Proc. Natl. Acad. Sci. USA* **102**, pp. 10421–10426, 2005.
7. V. R. Iyer *et al*, "The transcriptional program in the response of human fibroblasts to serum," *Science* **283**, pp. 83–87, 1999.
8. J. C. Gower and G. J. S. Ross *Appl. Stat.* **18**, pp. 54–64, 1969.
9. J. Eisner, *State-of-the-Art Algorithms for Minimal Spanning Trees*, University of Pensylvania, Department of Computer and Information Science, University of Pennsylvania, 1997.
10. R. Dobrin and P. Duxbury, "Minimum spanning trees on random networks," *Phys. Rev. Lett.* **86**, pp. 5076–5079, 2001.
11. G. J. Szabó, M. Alava, and J. Kertész, "Geometry of minimum spanning trees on scale-free networks," *Physica A* **330**, pp. 31–36, 2003.
12. D. Freifelder, *Molecular Biology*, Jones and Barlett Publishers, Inc., 40 Tall Pine Drive, Sudbury, MA 01776, 1983.
13. T. Aste, T. D. Matteo, and S. Hyde, "Complex networks on hyperbolic sufaces," *Physica A* **346**, pp. 20–26, 2005.
14. M. E. J. Newman, "Assortative mixing in networks," *Phys. Rev. Lett.* **89**, pp. 208701 1–46, 2002.
15. S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," *Physics Reports* **424**, pp. 175–308, 2006.
16. Y. Zhange *et al*, "Boolean networks using the chi-square test for inferring large-scale gene regulatory networks," *BMC Bioinformatics* **4062/2006**, pp. 402–407, 2006.
17. H. Kim, J. K. Lee, and T. Park, "Network motifs: Simple building blocks of complex networks," *Science* **8**, pp. 1471–2105, 2007.
18. K. Yugi *et al*, "A microarray data-based semi-kinetic method for predicting quantitative dynamics of genetic networks," *BMC Bioinformatics* **74**, pp. 48–94, 2002.